

Fast and straightforward analysis of charge transport data in single molecule junctions

Qian Zhang,^{a,b} Chenguang Liu,^c Shuhui Tao,^{a,b} Ruowei Yi,^{a,b} Weitao Su,^d Cezhou Zhao,^c Chun Zhao,^c Yannick J. Dappe,^e Richard J. Nichols^b and Li Yang^{*a,b}

Department of Chemistry, Xi'an-Jiaotong Liverpool University, Suzhou, 215123, China. E-mail: li.yang@xjtlu.edu.cn

Department of Chemistry, University of Liverpool, Liverpool, L69 7ZD, UK.

Department of Electrical and Electronic Engineering, Xi'an-Jiaotong Liverpool University, Suzhou, 215123, China.

College of Materials and Environmental Engineering, Hangzhou Dianzi University, 310018, Hangzhou, China.

SPEC, CEA, CNRS, Université Paris-Saclay, CEA Saclay 91191 Gif-sur-Yvette Cedex, France.

KEYWORDS. *Single molecule conductance, molecular electronics, scanning tunneling microscopy, data analysis.*

ABSTRACT: In this study, we introduce an efficient data sorting algorithm, including filters for noisy signals, conductance mapping for analyzing the most dominant conductance group and sub-population groups. The capacity of our data analysis process has also been corroborated on real experimental data sets of Au-1,6-hexanedithiol-Au and Au-1,8-octanedithiol-Au molecular junctions. The fully automated and unsupervised program requires less than one minute on a standard PC to sort the data and generate histograms. The resulting one-dimensional (1D) and two-dimensional (2D) log histograms give conductance values in good agreement with previous studies. Our algorithm is a straightforward, fast and user-friendly tool for single molecule charge transport data analysis. We also analyze the data in a form of a conductance map which can offer evidence for diversity in molecular conductance. The code for automatic data analysis is openly available, well-documented and ready to use, thereby offering a useful new tool for single molecule electronics.

Introduction

It is now possible to reliably measure the conductance of single molecules trapped between metallic electrodes. Important experimental factors here include achieving reliable and robust electronic and chemical coupling between the molecular bridge targets and the contacting electrodes.[1-3] Techniques capable of trapping single molecules between electrode contacts include the use of an scanning tunneling microscopy (STM),[4-6] conducting atomic force microscope (cAFM),[7] mechanically formed break junctions (MCBJ)[8] or nano-lithographically created gaps.[9] However, for all of these techniques, even the most reliable contact anchoring groups lead to a generally large variability in any given conductance determination.[10] This can be due to a number of factors including variable surface binding geometry between the molecule and contact, noise, environmental and contact fluctuations, multiple and interacting molecules in the gap and the stochastic nature of the junction formation and breaking processes.[11] It is then necessary to repeatedly record many traces and then statistically represent this data recorded for the making and breaking of many molecular junctions. In the simplest case, a single peak in a 1D conductance histogram resulting from plateau featured traces is usually taken to represent molecule junction formation and the most probable conductance value.[12-13] Such a statistical analysis has

been underpinning for achieving reliable electrical data for single molecule junctions.

There are then two broad approaches which can be applied to statistically analyse the data, each of which has its advantages and disadvantages. The first approach is to simply use all traces in the conductance histogram analysis. This has the advantage that there is no data preselection, however the analysis may include a lot of traces where either no or poorly contacted molecular junctions are formed. It may also be not possible to recognise molecular junction data if the probability of junction formation is low. The second approach is to try to select events where clear molecular junctions form from those where they do not. The advantage here is that data that is not relevant to molecular junction forming events is removed from the analysis, while molecular data is retained. The disadvantage here is that a preselection of data could introduce conscious or unconscious bias and affect the scientific interpretation of the determined conductance values.[14] It is of benefit to the single molecular electronics community to have a range of different analysis approaches available, including both the analysis of unselected data, as well as algorithmic approaches which attempt to recognise molecular junction forming events. The aim of the present study is to provide a new and well-documented approach to the latter.

Early work used hand-selected data through the recognition of plateau containing traces as an indicator of

good molecular junction formation, although it is recognised here that care has to be taken to avoid conscious or unconscious bias. On the other hand, a few algorithmic approaches to data selection have been presented in the literature. Jang *et al.* have introduced a last-step analysis (LSA) method for determining single molecule conductance, in which traces are selected which are believed to arise from molecular junction formation.[14] In this method, only curves featuring a very rapid conductance drop in the last step are selected. Gonzalez *et al.* presented an alternative method to extract the electrical conductance curves from a set of measured conductance traces.[15] Instead of selecting the plateau-containing curves, they consider all the traces and only remove background curves that feature a normal exponential decay of current synonymous with tunneling through an empty gap. Halbritter *et al.* introduced a two-dimensional correlation analysis of conductance traces to study the contact evolution in metal nano-junctions.[16] However automatic selection and analysis tools are also desirable for high throughput data analysis.

More recently an algorithmic method has been used for data sorting by Albrecht group.[17-18] Firstly for each exported current-distance trace the current signal is divided into many bins with a selected bin width (BW). Each bin will have a value for the number of plateaus-determining bin counts (PDBC). With a sorting algorithm which recognises plateau-containing traces from the bin counts, the traces containing plateau (molecular) features can be separated from both exponential traces and those primarily related to experimental noise.[17] Inspired by the widespread application of vector-based classification method in genetics, robotics and neuroscience,[19-20] they also developed an unsupervised vector-based classification of $I(s)$ traces which demonstrates how differently shaped $I(s)$ curves can be separated and grouped into clusters. Different event shapes are separated by this approach which essentially relies on a pattern analysis using several variables extracted for each current versus distance trace.[18]

The methods described above are based on the ideas of how to select plateau featured curves or how to group the $I(s)$ traces into clusters of traces showing similarities. In typical experimental sets instabilities in the molecular junction formation process, stochastic fluctuations in the microscopic details of the metal contacts or contamination of electrodes may all produce “noisy” traces which lead to the inclusion of traces in a data set which are not representative of proper molecular junction formation and consequently present difficulties in the subsequent data analysis. It is hence desirable to therefore effectively remove traces which do not show molecular junction formation.

Instead of such a targeted selection of plateau featured curves, we propose here a different approach to

data analysis, which does not select plateau featured traces, but removes undesired curves with clearly bad decay features as well as noisy traces, while keeping plateau-containing traces which are generally indicative of molecule junction formation. We demonstrate this method’s ability to treat actual experimental data obtained from the well-studied gold-octanedithiol-gold and gold-hexanedithiol-gold molecular junction systems.[21-22] The originality of this work lies in offering a new and straightforward scheme for automatic data analysis with good performance (the data processing time is less than one minute). The present approach can also offer evidence for diversity in molecular conductance.

Results and discussion

Data collection

The experiments were performed using the $I(s)$ technique, implemented according to the methodology described by Haiss *et al* [21] with necessary modification to our Bruker STM instrumentation. More detailed information can be also found in our previous works.[23-25] For the $I(s)$ measurements the gold STM tip was initially set at a given distance and then the tip was approached close to the substrate surface by adjusting the set-point current to high values. Current-distance traces were then recorded by rapid withdrawing the tip back to 4 nm at a rate of 7.66 nm/s. In the absence of molecular junction formation, an exponential decay of the current is observed (black curve in Figure 1b). However, upon close approach of the STM tip to the substrate molecules can also bridge the short gap through the stochastic formation of junctions. In such circumstance a plateau in the current-distance traces is typically observed, which is synonymous with the formation of a molecular junction (blue curve in Figure 1c). With further stretching of such molecular junctions the molecule becomes disconnected from the gold STM tip and a rapid drop of the current signal is then seen as a step at the end of the plateau (green curve in Figure 1d).

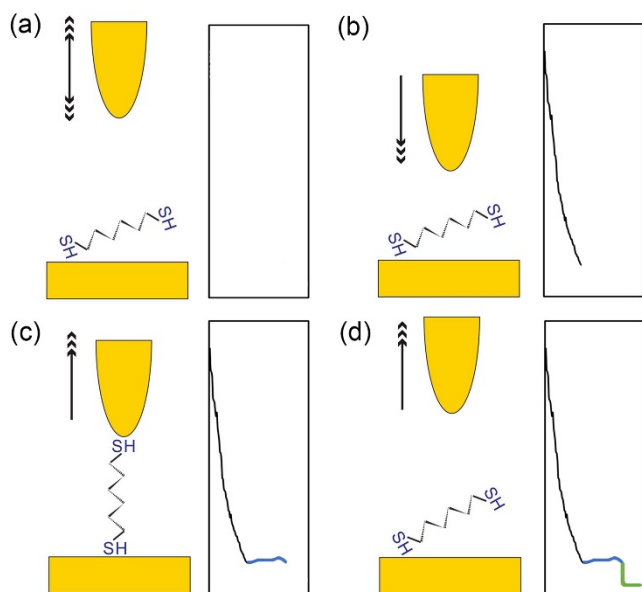


Figure 1. Schematic diagram of the $I(s)$ technique. (a) The initial stage of the molecular junction. (b) The gold tip is brought close to the gold substrate, the black curve represents the sharp decay of the current. (c) Formation of the Au-octanedithiol-Au molecular junction, the blue curve represents the plateau signal. (d) Breaking of the molecular junctions, with the green curve showing the rapid drop of the current.

Over 10000 current-distance curves were collected during this process with a preset bias voltage of 0.3 V and set-point current of 10 nA. The $I(s)$ traces can present highly diverse characteristics, such as pure decay (Figure 2, black), ideal plateau-featuring curves (Figure 2, blue), plateaus with noise spikes (Figure 2, green) and highly noisy curves (Figure 2, red). The preferred traces, taken as characteristic of stable molecular junction formation, are those featuring with the plateaus (blue), which signify defined formation of molecular junctions followed by their breaking. The aim here is then to sort the data to eliminate curves with abnormal shape or signs of excessive instability or noise.

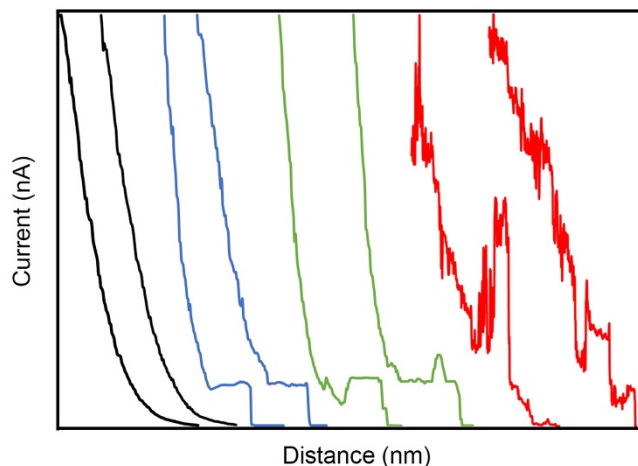


Figure 2. Typical $I(s)$ traces experimentally obtained (black, without molecular junction formation; blue, ideal plateau-featuring curves; green, plateau-featuring curves with noisy peaks; red, very noisy curves).

Generally, two main analysis pathways can be applied to these exported data (Figure 3). In the conventional method, the collected $I(s)$ data is plotted in a visualized way as 1D or 2D conductance (junction separation distance) histograms, with plateau featuring curves being selected for the data visualization. However, this approach relies on the signal shape and therefore can impose an assumptive outcome on the final histograms. To avoid this issue, consistent criteria for plateau selection should be applied for every data set, ideally with data analysis or selection which does not make any overly restrictive a priori assumptions about the data shape. In our data sorting routine, the $I(s)$ data is sent through a series of filters to remove any noisy signals caused by, for instance, contamination of the electrodes, passivation of the STM tip, fluctuations of the molecular junctions and so on. The follow filters are applied in a stepwise manner: X-Filter, Y-Filter, Peak-Filter. Finally, the conductance mapping distributes the majority of the plateau featured traces and the sub-populations in one histogram.

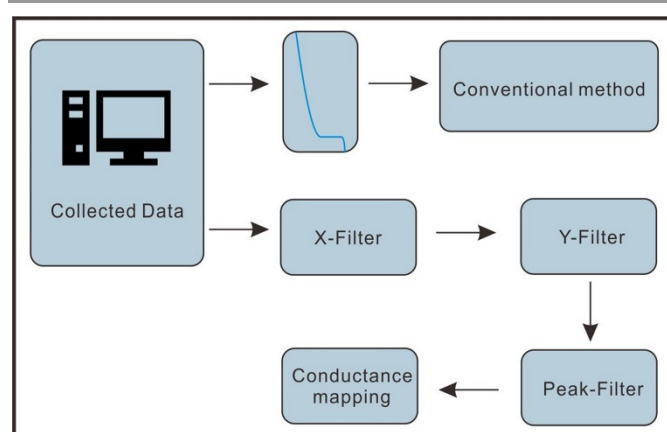


Figure 3. Flowchart of two main data analysis pathways. One is the conventional method of plotting the $I(s)$ signal in a visualized way using the conventional 1D or 2D histograms. The other one is our automatic data sorting algorithm, which includes an X-Filter, Y-Filter, Peak-Filter and conductance mapping functions, as explained in the text.

X-Filter

Each $I(s)$ file consists of distance and current signals in m rows as shown in Figure 4a, each row of the data is a single point in the plotted $I(s)$ curve. For example, the first point $p_1 (d_1, c_1)$, the second point $p_2 (d_2, c_2)$, the third one $p_3 (d_3, c_3)$ and so on, aggregate to a total number of N rows resulting in a data matrix. This matrix is the fundamental platform for further processing. Ideally in the absence of molecular junction formation the current will decay exponentially as a function of the increasing distance between the gold tip and substrate, and a well decaying current should approach 0 nA as the distance moves towards the limit of the tip retraction. Also in the presence of molecular junction formation the current should also decay to zero as the molecular bridge is broken and the tip is moved to the limit of its retraction. However, in the real experiment, a certain fraction of retraction traces may show an abnormal decay of current. In these cases the current does not proceed effectively towards zero or there are noise features or peaks which appear in the low current range. Our X-Filter algorithm aims at removing these noisy curves caused by the abnormal decay of the current. Here, we consider the distance signal as the X-axis, and the current signal as the Y-axis. For values above 2 nm ($d_n > 2$) in the X-axis, the corresponding values in Y-axis have been used for calculating the mean value (μ) and variance (σ^2) (Figure 4b). The equations for mean value (1) and variance (2) are simple but quite effective.

$$\mu = (c_1 + c_2 + c_3 \dots + c_N) / N \quad (1)$$

$$\sigma^2 = (c_1 - \mu)^2 + (c_2 - \mu)^2 + \dots + (c_N - \mu)^2 / N \quad (2)$$

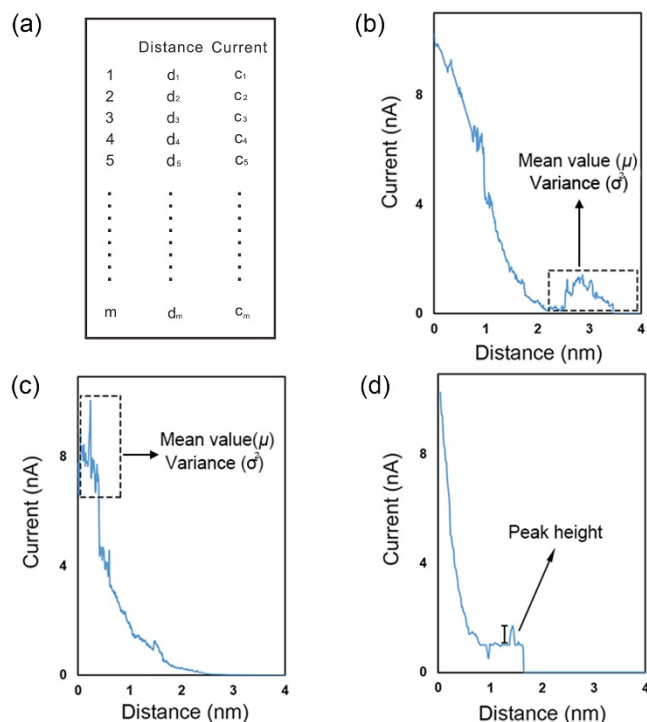


Figure 4. (a) Structure of the $I(s)$ data and schematic diagram of X-Filter (b), Y-Filter (c) and Peak-Filter (d) functions.

The mean value is used to check whether the current decays to 0 nA, while the variance is used to express the quality of the data being different, divergent or inconsistent. Using a given threshold of mean value and variance, for example, $\mu_c < 0.1$, $\sigma_c^2 < 0.1$, only decay curves close to normal would follow this criterion. The chosen mean value and variance as a setting criterion relies on the trial and error scheme (see detailed explanation of the scheme and parameters included in SI). The trial and error values can be adjusted many times, with the result being checked in the exported figures. Here, we use 0.1 as the criterion to make sure the current signals decay to close to 0. As a result, abnormally decaying $I(s)$ curves are removed by the X-Filter.

Y-Filter

The Y-Filter is also based on the mean value and variance scheme, but here there is a focus on the very beginning of the current decay. When the gold STM tip reaches the preset tunneling region, which is 10 nA in our Au-1,8-octanedithiol-Au system, the current should drop sharply with the retraction of the tip during the very initial stages of junction opening. Some $I(s)$ traces exhibit an oscillating or excessively noisy signal at the beginning of the current decay which may arise from poorly contacted molecules, contaminated tips or multiple interacting molecules in the junction. It is beneficial to eliminate such curves from the final conductance histograms. By locating the data points at the high range of the current, the mean value of the corresponding

distance and the variance of these current values have been calculated. For curves which decay well in the initial stages of the retraction corresponding to the current above 6 nA ($y > 6$), the variance of the current should be small (i.e. close to the y axis). There should be a rapid decay of this current and the mean value of the corresponding distance should be close to zero ($\mu_d \sim 0$). However, in the case of Figure 4c, the noisy signal would both bring a higher value of the mean of distance (μ_d) and the current variance (σ^2_c). As a consequence, the mean value of the distance and the variance of the current have been used to define the decay quality in the high current range. To find the optimum mean and variance values for data filtration, the values obtained from well decaying curves can be used as a reference.

Peak-Filter

Subsequently, we investigated the noisy peaks present in the plateau region. For $I(s)$ traces with otherwise reasonable current decay forms, the linear plateau can show oscillating signals (small peaks) as shown in Figure 4d. These traces may also be deemed as undesired curves, and if there are sufficient of them they could lead to broadening or shoulder peaks in the conductance histograms. By comparing the mathematical difference of the current values, the location, width and height of the peaks can be obtained. A local peak is a data sample that is larger than its two neighbouring samples with the peak height taken as the difference between the peak data and its neighbouring samples. If we defined seven current points (1, 2, 1, 3, 1, 5, 1) as a simple illustrative example, then three local peaks (2, 3, 5) with peak height (1, 2, 4) are calculated above the plateau current of 1. Following this regime, the number of the local peaks and the peak height can be then used to classify the data. Although, the $I(s)$ data can present diverse shapes, such as the slanted plateau and non-linear with high or low amplitude, the Peak-Filter focuses on the removal of telegraphic noise on the plateau, the other plateau traces being then sent to the conductance mapping process.

the ideal plateau region, a dominant plateau group indicates the most probable conductance value of the molecular junction.

Conductance mapping

Finally, we use a conductance mapping process to obtain the most dominant conductance value for a given molecular target. For a single $I(s)$ trace constructed from 512 data points, the Y-axis (current axis) was simply divided into many steps (bins). In each bin, the counts of the data points are different and the plateau region will always have a larger bin count than those adjacent bins. We have also noticed that, since the current decays to zero, the first bin will give very large counts, and therefore we excluded this bin to eliminate the effect of the high count value for the first bin. As shown in figure 5a, the current value was divided into 15 bins (bin width = 0.7 nA), the plateau was included in the second bin with a bin counts of 50, and the other bins possess a smaller bin counts of 10. Instead of simply using the bin counts as a criterion to separate the pure decay curve and plateau featured curves which is described by Inkpen *et al.*,^[17] here we introduce a new approach for analysing each population of the data set. We select the maximum bin counts to calculate the current mean value of these data points. For example, the maximum counts in Figure 5a is 50, and consequently we calculate the mean value of these 50 data points to get a plateau mean value of 1.1 nA. Besides the plateau counts, a conductance mapping has been plotted as shown in Figure 5b. In this map, the $I(s)$ curves have been transformed as blue points according to their plateau mean values (X-axis) and plateau counts (Y-axis). The conductance map was then grouped in three main parts based on the plateau counts. For a region with plateau counts lower than 15 (black rectangular region) in Figure 5b, we deemed it as the decay region, including the pure decay featuring $I(s)$ curves. The point counts for a pure decay curve is low and hence this region lies at the bottom of the plot. Similarly, we treat the plateau counts above 15 but lower than 30 as the short plateau region (green rectangular region) and the plateau counts above 30 as the ideal plateau region (blue rectangular region). In the ideal plateau region, we find that the distribution of blue points has a dominant region (red oval region) and a less dominant region (purple oval region). Even though the plateau position for these featured curves is different, we do observe a dominant region of the plateau mean value to indicate the most probable conductance value for the molecular junction. Notably, for a particular molecular junction, we can also observe multiple plateaus in the $I(s)$ curves and low or high conductance peaks in the 1D histogram.^[22,26] This is ascribed either to the formation of multi-molecule junctions or to different molecular junction configurations.^[7,27] It is thus possible to analyze if the molecular junction is a dominant conformation scheme (one dominant region) or if it has a

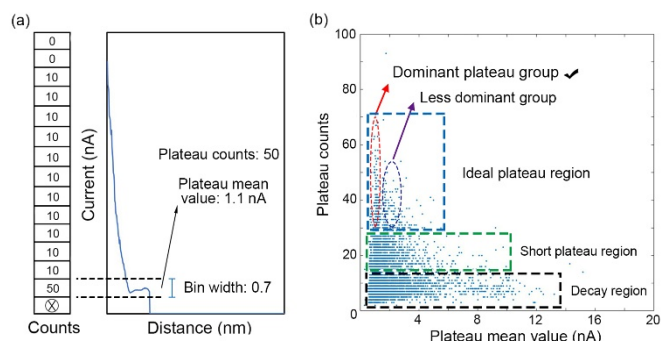


Figure 5. (a) The counts scheme for obtaining the plateau mean value and plateau location. (b) Conductance mapping histogram, including the decay region, short plateau region and ideal plateau region. In

multiple peak scheme. It is noteworthy that the parameters and settings included in the conductance mapping do not rely on the subjective choices of the operator, but are based on a trial and error scheme (see SI). An appropriate setting is achieved after obtaining clear 1D and 2D histograms. Notably, the software can list and try many possible bin width and bin count combinations in a very short time.

Experimental results

We first apply the noise removal and conductance mapping algorithm to the experimental data for Au-1,8-octanedithiol-Au molecular junctions. Over 11,000 $I(s)$ traces were collected at a bias voltage of 0.3 V, a current set point of 10 nA and a tip displacement of 4 nm from the initial separation at the set-point current. The raw data were directly exported as ASCII text files without any treatment and then analysed by the algorithm.

In Figure 6a, the conductance map was firstly grouped to three main regions based on the plateau counts. As discussed above, region 1 (green) was deemed to correspond to the decay region without any plateau featured curves included, region 2 (red) reflected the $I(s)$ curves with short plateau or small noisy peaks, and the region 3 (blue) related to the plateau region with linear plateau curves of sufficient extension. To gain insights in these plateau featuring curves, the plateau region was then sorted in the horizontal direction based on the plateau mean value (position of the plateau), and five refined regions were then classified, as shown in Figure 6b. One can clearly observe that region 5 (red) presents a dominant number of data points, indicating the highest probability of the plateau position. The $I(s)$ curves in region 5 were then plotted as 1D and 2D histograms.

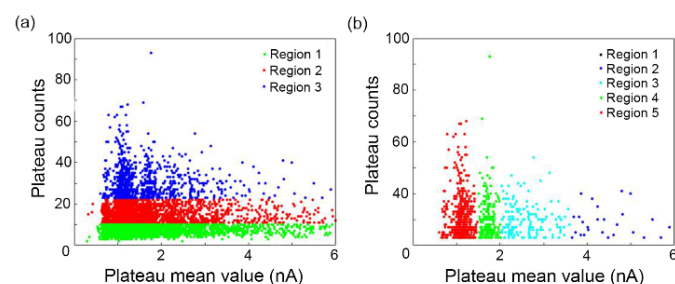


Figure 6. (a) Conductance mapping for Au-1,8-octanedithiol-Au molecular junctions grouped in three main regions based on the plateau counts, (b) Conductance mapping grouped in five refined regions to get insights in the most dominant conductance peak.

In the 1D conductance histogram, a dominant peak located at 3.67 nS ($4.75 \times 10^{-5} G_0$) is observed. The corresponding yellow area around this conductance value in the 2D log histogram indicates the distribution of the plateau featuring traces (Figure 7a,b). This conductance value agrees well with previously reported

values for Au-1,8-octanedithiol-Au molecular junctions which shows that our automatic data analysis approach is reliability performing in this test.[6,14,28] For comparison, 1D and 2D histograms are also plotted in Figure 7c and 7d for the raw data without the use of the algorithm. Since these unfiltered histograms include a lot of traces where either no or poorly contacted molecular junctions form, the conductance peak was then hidden by the large proportion of deleterious background curves. There was a quandary at the early stages of the single molecule conductance technique development that different values of conductance were measured using different methods.[29-31] For the break junction method (BJ), both high and low current steps were found, with a primary conductance value of 20 nS being obtained from the BJ technique for gold-1,8-octanedithiol-gold junctions.[2] On the other hand, Haiss measured a primary conductance value of 1 nS using the $I(s)$ method.[21] Nichols *et al.* suggested this observation could be related to the coordination differences of the sulfur anchoring group to the gold electrodes.[27] Three difference conductance values were attributed to different contact configurations of the thiolate binding to gold surface atoms. Our experimental data found using the $I(s)$ method and conductance mapping process verifies such a conductance diversity and the existence of differing conductance groups. In the conductance mapping, a dominant red region was observed, indicating the most abundant class of plateau featured traces, the other regions being sub-populations caused by other configurations.

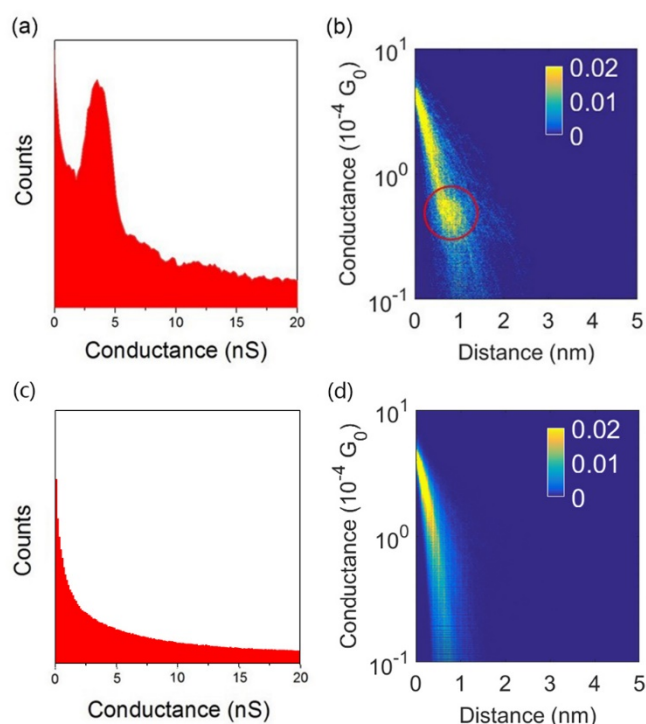


Figure 7. (a) 1D histograms of single-molecule conductance of gold-1,8-octanedithiol-gold junctions, a dominant peak was located at 3.67 nS. (b) The corresponding 2D log histogram with a sensitivity indicator of the conductance counts. 1D (c) and 2D (d) log histogram of all the raw $I(s)$ data without the treatment of the algorithm.

Our algorithm has also been applied to 1,6-hexanedithiol, following a similar procedure to that just described for 1,8-octanedithiol. The resulting conductance map in Figure 8a, was grouped in three main regions, including the decay region (green), short plateau region (red) and the ideal plateau region (blue). The data points at ideal plateau region were then plotted as a refined conductance map with five colour coded regions marked. In this map, the light blue region dominates, indicating the most probable conductance value of the Au-1,6-hexanedithiol-Au molecular junction. It is also worth noting here a less dominant red region also appears on this map, representing the possibility of forming a molecular junction with a smaller conductance value. We treat the dominant region (light blue region) as the main conductance value of 1,6-hexanedithiol molecular junctions and the related data was then plotted as 1D and 2D log histograms.

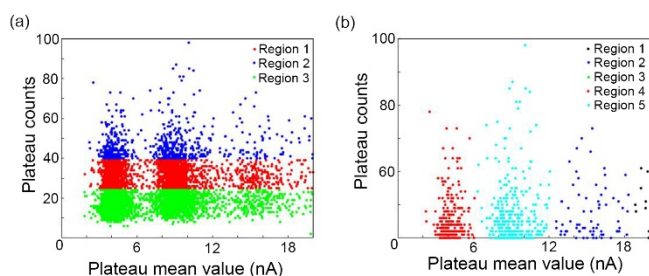


Figure 8. (a) Conductance mapping for Au-1,6-hexanedithiol-Au molecular junctions with colour coded grouping into three main regions based on the plateau counts. (b) Conductance mapping of the ideal plateau region, five refined regions were grouped to examine the most dominant conductance peak and sub-groups.

Figure 9a represented the 1D conductance histogram of selected plateau traces, a dominant peak located at 29 nS ($3.75 \times 10^{-4} G_0$) is observed, the corresponding yellow area in 2D log histogram indicated the distribution of the plateau featuring traces. The 1D and 2D histograms of all curves without any algorithmic analysis have also been plotted (Figure 9c,d). In comparison with the previously reported value (28 nS), our data agrees very well with the literature data.[6,14] We also noticed that the less dominant region (red in Figure 8b) around 4.3 nA showed a respectable number of plateau featured curves, and hence our algorithm can be used to visually assess the number and distribution of each population. The most dominant region (light blue in Figure 8b) in conductance mapping was considered then to correspond to the most

probable configuration of Au-1,6-hexanedithiol-Au molecular junctions with a conductance value of 29 nS.

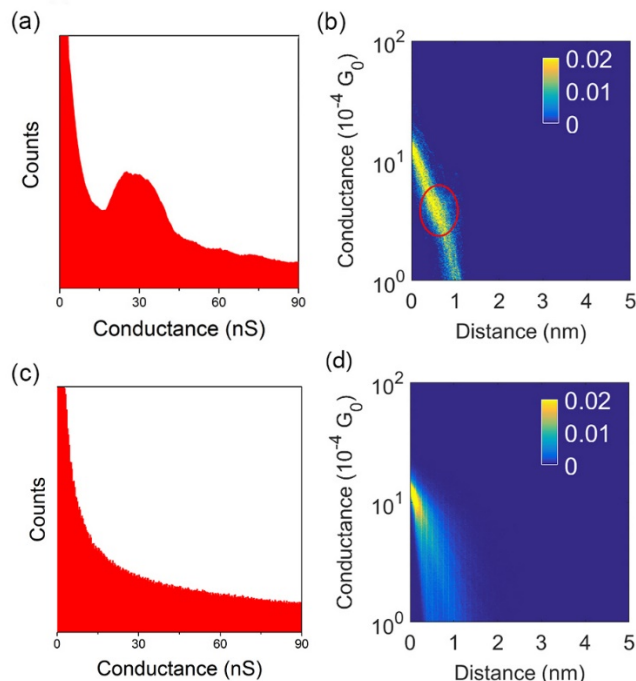


Figure 9. (a) 1D histograms of single-molecule conductance of gold-1,6-hexanedithiol-gold junctions, a dominant peak was located at 29 nS. (b) The corresponding 2D log histogram with a sensitivity indicator of the conductance counts, a higher density yellow region was located at $3.75 \times 10^{-4} G_0$. 1D (c) and 2D (d) log histogram of all the raw $I(s)$ data without the treatment of the algorithm.

Selection of appropriate parameter filters is important for an effective implementation of the algorithm. Here, we take the data from 1,6-hexanedithiol molecular junctions as the example to demonstrate the capabilities of each filter. Please note that the parameters chosen do not simply rely on the subjective choices of the operator, but are based on the trial and error scheme (see detailed explanation in SI). Values can be adjusted many times until the clear Gaussian peaks are observed in the exported histograms.

In Figure 10, we performed some tests to demonstrate the effect of different algorithm parameters. The 1D histogram in Figure 10a was deemed to represent the preferred implementation as it gives a well-defined conductance peak, indicating the most probable conductance value of the molecular junction. This histogram was plotted setting the distance range of 0-2.8 nm for the X-Filter, a variance of 30 for Y-Filter, and a value of 3 for the Peak-Filter. As discussed earlier, the X-Filter algorithm is aimed at removing noisy curves caused by the abnormal decay of the current, by given a threshold of the mean value and variance for the selected distance range, for example, $d=0-2.8$ nm $\mu < 0.1$, $\sigma^2 < 0.1$.

In Figure 10b, the distance range of the X-Filter was decreased to 0-2 nm, which means more curves could comply with this criterion and consequently contribute noise from the initial decay region. Compared with Figure 10a, only the variance value for the Y-Filter was changed to 8 in Figure 10c, indicating a more stringent criterion. Since the Y-Filter focuses on undesired oscillating or noisy signals at the very beginning of the current decay, the lower variance value cleans up the histogram. Figure 10d demonstrates the efficiency of the Peak-Filter, by changing the tolerance of peak numbers from 3 to 6, more plateau featuring curves with oscillating signals (small peaks) on the plateau could fit the criterion, the resulting shoulder peaks were then observed. These results demonstrate that some adjustments of the algorithm parameters are necessary for each molecular junction, but even with these adjustments the main peak conductance remains broadly unchanged

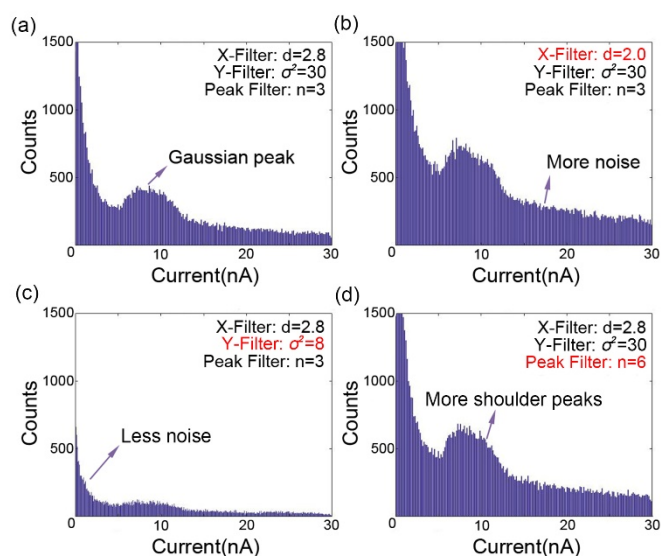


Figure 10. (a) 1D histogram of Au-1,6-hexanedithiol-Au molecular junctions constructed using the given filters. These are an X-Filter at the distance of 2.8 ($d=2.8$), the variance of the Y-Filter was set to 30 ($\sigma^2=30$), the number of peaks for the Peak-Filter was set to 3 ($n=3$). (b) 1D histogram plotted using the parameters of X-Filter: $d=2.0$, Y-Filter: $\sigma^2=30$, Peak-Filter: $n=3$. (c) 1D histogram after the application of filter using the following parameters: X-Filter: $d=2.8$, Y-Filter: $\sigma^2=8$, Peak-Filter: $n=3$. (d) 1D histogram using an X-filter of $d=2.8$, Y-Filter of $\sigma^2=30$, Peak-Filter of $n=6$.

Experimental

The experimental data sets for gold-1,8-octanedithiol-gold molecular junction were collected by the STM based $I(s)$ technique. Gold coated substrate was provided by Arrandee (Germany) with a further in-house flame annealing process. 1,6-hexanedithiol, 1,8-octandithiol and mesitylene were bought from Alfa Aesar and used as received and the distilled water was supplied by the

central purification system in our chemical research laboratory. Gold substrates were functionalized with either hexanedithiol or octanedithiol in mesitylene solutions (10 mM) in a liquid cell. Gold wire (99.99%) with 0.25 mm diameter was purchased from Tianjing Lucheng Metal and then etched in a solution electrolyte of hydrochloric acid and ethanol (50:50, v:v). Electrochemical tip etching was performed using the method described by Ren *et al.*,[32] and the STM tips were then characterized by scanning electron microscopy (SEM) to ensure a good shape of the etched tip. The $I(s)$ technique used here is described in the data collection section and is also elaborated in more detail in our previous studies.[23-25] The data analysis algorithm was written in Matlab (2016b) and the detailed classification approach is described in the main text. All relevant scripts and data can be found in SI and are also available from the authors on request. When setting the parameters of specific data sets, several attempts can be made to find the optimum values. The detailed parameters used in this work are shown as table 1 below. Using the optimum parameters for the code, the most dominant conductance cluster has been located after the data sorting program.

Table 1. Parameters used in the experiment.

	1,6-hexanedithiol	1,8-octanedithiol
X-Filter	Data range: 0-2.8nm Mean value: 0.1 Variance: 0.1	Data range: 0-2nm Mean value: 0.1 Variance: 0.1
Y-Filter	Data range: 20-30nA Mean value: 3.5 Variance: 30	Data range: 6.6-10nA Mean value: 3.5 Variance: 10
Peak Filter	Peak number: 3	Peak number: 3
Conductance mapping	Number of steps: 5	Number of steps: 19

Conclusions

We present here a simple, fast and user-friendly algorithm for the analysis of single molecule conductance data obtained by the $I(s)$ STM method. This could readily be extended to other methods such as the STM-BJ technique or mechanically controlled break junctions. In the supporting information, we provide fully documented MATLAB routines to promote usage of this analysis package by the molecular electronics community. The automatic data analysis process focuses on the removal of traces where proper molecule junction formation does not occur, such as excessively noisy traces as the molecular bridge is formed or traces that do not properly decay to zero current when the molecular junction is cleaved. Through the application of noise removal algorithms, the difficulties caused by instability,

fluctuations and noise in single molecule junction formation have been reduced, and data pertaining to effective junction formation can be separated from interfering signals. The remaining traces are then displayed using a conductance mapping process which makes it possible to analyse the statistical diversity in the dataset and recognise groups and sub-groups in the data. This captures the statistical complexity of the molecular system in a straightforward way enabling analysis of the effects of molecular structure, contacts, environment and other physical parameters. The algorithms we present here are also amenable to the analysis of very large datasets. The algorithm is a simple and direct method in which the raw *I(s)* text files are simply copied into the same folder as the program file, the parameters are set and code is simply launched to run the analysis. For experimental groups having a different text structure to ours, only one extra step is needed to convert the text data files. We have applied the algorithm to our experimental data for Au-1,8-octanedithiol-Au and Au-1,6-hexanedithiol-Au molecular junctions. Following the proscribed analysis, a clear conductance peak is seen in the histograms and the conductance values are in good agreement with previous studies. Future development could include coding which eliminates the need to manually set parameters with the algorithm finding suitable parameters itself, for example by exploiting machine learning algorithms.

Conflicts of interest

The authors declare no competing financial interest.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC Grants 21503169, 2175011441), the Jiangsu Science and Technology program (BK 20140405), Suzhou Industrial Park Initiative Platform Development for Suzhou Municipal Key Lab for New Energy Technology (RR0140) and the XJTLU Research Development Fund (PGRS-13-01-03 and RDF-14-02-42).

References

- 1 A. Aviram and M. A. Ratner, *Chem. Phys. Lett.*, 1974, **29**, 277-283.
- 2 B. Xu and N. J. Tao, *Science*, 2003, **301**, 1221-1223.
- 3 Y. Cao, S. Dong, S. Liu, L. He, L. Gan, X. Yu, M. L. Steigerwald, X. Wu, Z. Liu and X. Guo, *Angew. Chem. Int. Ed.*, 2012, **124**, 12394-12398.
- 4 W. Haiss, H. van Zalinge, S. J. Higgins, D. Bethell, H. Höbenreich, D. J. Schiffrin and R. J. Nichols, *J. Am. Chem. Soc.*, 2003, **125**, 15294-15295.
- 5 L. Venkataraman, J. E. Klare, I. W. Tam, C. Nuckolls, M. S. Hybertsen and M. L. Steigerwald, *Nano Lett.*, 2006, **6**, 458-462.
- 6 F. Chen, X. Li, J. Hihath, Z. Huang and N. Tao, *J. Am. Chem. Soc.*, 2006, **128**, 15874-15881.
- 7 X. D. Cui, A. Primak, X. Zarate, J. Tomfohr, O. F. Sankey, A. L. Moore, T. A. Moore, D. Gust, G. Harris and S. M. Lindsay, *Science*, 2001, **294**, 571-574.
- 8 M. A. Reed, C. Zhou, C. J. Muller, T. P. Burgin and J. M. Tour, *Science*, 1997, **278**, 252-254.
- 9 X. Guo, J. P. Small, J. E. Klare, Y. Wang, M. S. Purewal, I. W. Tam, B. H. Hong, R. Caldwell, L. Huang, S. O'Brien, J. Yan, R. Breslow, S. J. Wind, J. Hone, P. Kim and C. Nuckolls, *Science*, 2006, **311**, 356-359.
- 10 M. Ratner, *Nat. Nanotech.*, 2013, **8**, 378.
- 11 R. Frisenda, M. L. Perrin, H. Valkenier, J. C. Hummelen and H. S. J. van der Zant, *J. Phys. Status Solidi B*, 2013, **250**, 2431-2436.
- 12 W. Haiss, C. Wang, I. Grace, A. S. Batsanov, D. J. Schiffrin, S. J. Higgins, M. R. Bryce, C. J. Lambert and R. J. Nichols, *Nat. Mater.*, 2006, **5**, 995.
- 13 C. Wang, A. S. Batsanov, M. R. Bryce, S. Martín, R. J. Nichols, S. J. Higgins, V. M. García-Suárez and C. J. Lambert, *J. Am. Chem. Soc.*, 2009, **131**, 15647-15654.
- 14 S. Y. Jang, P. Reddy, A. Majumdar and R. A. Segalman, *Nano Lett.*, 2006, **6**, 2362-2367.
- 15 M. T. González, S. Wu, R. Huber, S. J. van der Molen, C. Schönenberger and M. Calame, *Nano Lett.*, 2006, **6**, 2238-2242.
- 16 A. Halbritter, P. Makk, S. Mackowiak, S. Csonka, M. Wawrzyniak and J. Martinek, *Phys. Rev. Lett.*, 2010, **105**, 266805.
- 17 M. S. Inkpen, M. Lemmer, N. Fitzpatrick, D. C. Milan, R. J. Nichols, N. J. Long and T. Albrecht, *J. Am. Chem. Soc.*, 2015, **137**, 9971-9981.
- 18 M. Lemmer, M. S. Inkpen, K. Kornysheva, N. J. Long and T. Albrecht, *Nat. Commun.*, 2016, **7**, 12922.
- 19 J. V. Haxby, A. C. Connolly and J. S. Guntupalli, *Annu. Rev. Neurosci.*, 2014, **37**, 435-456.
- 20 T. Naselaris, K. N. Kay, S. Nishimoto and J. L. Gallant, *NeuroImage*, 2011, **56**, 400-410.
- 21 W. Haiss, R. J. Nichols, H. van Zalinge, S. J. Higgins, D. Bethell and D. J. Schiffrin, *Phys. Chem. Chem. Phys.*, 2004, **6**, 4330-4337.
- 22 X. Li, J. He, J. Hihath, B. Xu, S. M. Lindsay and N. Tao, *J. Am. Chem. Soc.*, 2006, **128**, 2135-2141.
- 23 Q. Zhang, L. Liu, S. Tao, C. Wang, C. Zhao, C. González, Y. J. Dappe, R. J. Nichols and L. Yang, *Nano Lett.*, 2016, **16**, 6534-6540.
- 24 Q. Zhang, S. Tao, R. Yi, C. He, C. Zhao, W. Su, A. Smogunov, Y. J. Dappe, R. J. Nichols and L. Yang, *J. Phys. Chem. Lett.*, 2017, **8**, 5987-5992.
- 25 L. Liu, Q. Zhang, S. Tao, C. Zhao, E. Almutib, Q. Al-Galiby, S. W. D. Bailey, I. Grace, C. J. Lambert, J. Du and L. Yang, *Nanoscale*, 2016, **8**, 14507-14513.
- 26 X. S. Zhou, Z. B. Chen, S. H. Liu, S. Jin, L. Liu, H. M. Zhang, Z. X. Xie, Y. B. Jiang and B. W. Mao, *J. Phys. Chem. C*, 2008, **112**, 3935-3940.
- 27 R. J. Nichols, W. Haiss, S. J. Higgins, E. Leary, S. Martin and D. Bethell, *Phys. Chem. Chem. Phys.*, 2010, **12**, 2801-2815.
- 28 T. Morita and S. Lindsay, *J. Am. Chem. Soc.*, 2007, **129**, 7262-7263.
- 29 J. L. Xia, I. Diez-Perez and N. J. Tao, *Nano Lett.*, 2008, **8**, 1960-1964.
- 30 C. Li, I. Pobelov, T. Wandlowski, A. Bagrets, A. Arnold and F. Evers, *J. Am. Chem. Soc.*, 2008, **130**, 318-326.
- 31 M. Fujihira, M. Suzuki, S. Fujii and A. Nishikawa, *Phys. Chem. Chem. Phys.*, 2006, **8**, 3876-3884.
- 32 B. Ren, G. Picardi and B. Pettinger, *Rev. Sci. Instrum.*, 2004, **75**, 837-841.